Multimodal Data Analysis: Open Science, Ethics, Archiving via DMPs

EnvisionBOX: Computational Reproducibility and Ethical Data Management

Babajide Owoyele and Wim Pouw

Winter School 2025 - Tuesday Session

Multimodal Data Analysis: Open Science, Ethics, Archiving via DMPs

EnvisionBOX: Computational Reproducibility and Ethical Data Management

Babajide Owoyele and Wim Pouw

Winter School 2025 - Tuesday Session

Tuesday Overview: Open Science & Data Ethics

Today's Focus

- Quick Intro to EnvisionBOX and Recap
- Open Science Principles
- Computational Reproducibility
- Ethical Data Archiving
- Masking Techniques

Schedule

1. Conceptual Session (15 min)
2. Q&A (5 min)
3. Technical Session: Masking 101 (20 min)
4. Collaborative Discussion on DMPs and Masking Tensions (30 min)

Open Science in Multimodal Research

Multimodal Research Challenges

- Complex data collection
- Diverse analysis techniques
- Interdisciplinary methodologies
- Ethical data management

Open Science Principles

Core Values:

- Transparency in methods vs Data accessibility
- Computational reproducibility vs Collaborative knowledge sharing

Practical Implementation:

- Shared data repositories
- Documented analysis pipelines
- Open-source tools
- Ethical data masking

Why Open Science Matters in Multimodal Research

Computational Aspects

- Standardize analysis methods
- Enable cross-study comparisons
- Reduce methodological variations
- Support peer verification

Community Benefits

- Lower entry barriers
- Accelerate research progress
- Promote interdisciplinary collaboration
- Enhance research integrity

Key Principle

Open as possible, closed as necessary: Balancing data accessibility with ethical considerations

Open Science Principles

Core Values

- Transparency
- Accessibility
- Reproducibility
- Collaboration

Implementation

- Open Data
- Open Methods
- Open Source Code
- Open Access
- OPEN EXCHANGE

Building Open eXchange (BOX)

Community of Learners

- Promote method literacy beyond result reporting
- Welcome contributions from all researchers
- Tailor communication to guide method reproduction

Didactical Approach

- Modules with practical instructions
- Introduce concepts and routines
- Focus on learning, not just code sharing

Building Open eXchange (BOX): Key Principles

Self-Ownership

- Proper citation for contributors
- Acknowledge method creators
- Recognize intellectual contributions

Build to Grow

- Expand platform scope
- Host theoretical frameworks
- Curate research updates
- Community-driven development

Community Contribution

- Interested in helping? Reach out and share your ideas!

Computational Reproducibility

Key Elements

- Version Control (Git)
- Environment Management
- Documentation
- Data Provenance

Best Practices

- Clear Directory Structure
- Requirements Documentation
- Code Comments
- README Files

Computational Reproducibility: Foundations

What is Computational Reproducibility?

- Ability to recreate computational results
- Transparency in research methods
- Consistent outcomes across different environments
- Verification of scientific claims

Why Matters in Linguistics

- Validate novel language analysis techniques
- Ensure reliability of computational methods
- Support collaborative research
- Enhance research credibility

Linguistic Research Reproducibility Challenges

Data Challenges

- Diverse data sources
- Multilingual corpora
- Variability in language samples
- Contextual nuances

Methodological Hurdles

- Complex preprocessing
- Annotation inconsistencies
- Software version dependencies
- Computational environment variations

Key Reproducibility Risks

- Inconsistent data cleaning
- Undocumented preprocessing steps
- Lack of version control

Practical Reproducibility Toolkit

Version Control Essentials

- Git repositories
- GitHub/GitLab for project management
- Commit documentation
- Branch management for experimental approaches

Computational Tools

- Jupyter Notebooks
- Conda environments
- Requirements.txt files
- Virtual environments

Documentation

- Detailed README files
- Code comments
- Method descriptions
- Dependency tracking

Reproducibility in Linguistic Analysis

Recommended Workflow

1. Data Collection
2. Preprocessing Documentation
3. Computational Pipeline
4. Analysis Verification
5. Result Archiving

Sharing Best Practices

- Use open-source platforms
- Provide complete analysis scripts
- Include sample datasets
- Document environment specifications

Practical Reproducibility Scenarios

Research Master Perspective
Scenario: Analyzing Multilingual Corpus

- Collect language samples
- Develop computational analysis method
- Ensure method can be replicated by peers

Reproducibility Checklist

- Annotate all preprocessing steps
- Share complete computational environment
- Provide clear method explanation
- Allow external verification

Overview

- Why Reproducibility Matters
- Version Control Fundamentals
- Environment Management
- Documentation Best Practices
- Data Management
- Common Pitfalls

The Reproducibility Crisis

Common Scenarios

- Analysis works today, breaks tomorrow
- Code runs on your machine only
- Data processing steps forgotten

Impact

- Time wasted
- Research blocked
- Trust diminished

Key Question

How can we make research that stands the test of time?

Version Control: Git Basics

Essential Commands

```
# Initialize repository
git init

# Track changes
git add .
git commit -m "Meaningful message"

# Share changes
git push origin main
```

Benefits

- Track changes
- Collaborate easily
- Backup work

Environment Management

Virtual Environments

```
# Create virtual environment
python -m venv myenv

# Activate
source myenv/bin/activate  # Unix
myenv\Scripts\activate     # Windows

# Save dependencies
pip freeze > requirements.txt
```

Why It Matters

Ensures code runs the same way everywhere

Documentation Hierarchy

Project Level

- README.md
- Requirements
- Project structure
- Installation guide

</> Code Level

- Function docstrings
- Comments
- Type hints
- Usage examples

Data Management

File Organization

```
project/
 data/
    raw/
    processed/
 src/
    preprocessing/
    analysis/
 results/
 docs/
```

Best Practices

- Use relative paths
- Version control data when feasible

Common Pitfalls

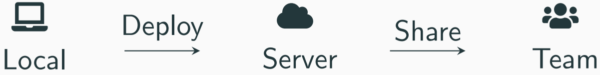
Environment Issues

- Undocumented dependencies
- Hard-coded paths
- Missing environment files

Version Control Mistakes

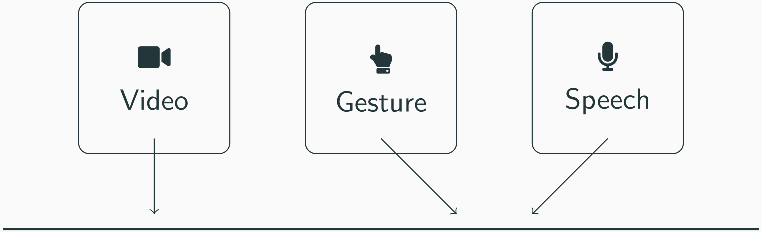
- Large files in Git
- Sensitive data exposed
- Poor commit messages

The "Works on My Machine" Problem



Solution
Containerization ensures consistent environments

The "Time Alignment" Problem



Problem
Different software tools = Different time references

Solution

- Establish sync points
- Use consistent time codes
- Document time transformations

The "Lost Metadata" Nightmare

Initial Collection

- Demographics
- Language profiles
- Experience data

6 Months Later

- Which was which?
- Who was bilingual?
- Missing context

Solution

Create standardized metadata templates

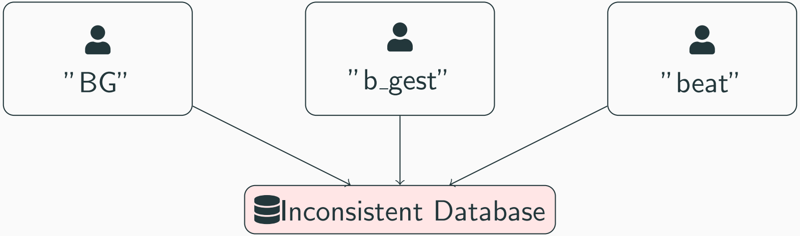
The "Format Compatibility" Crisis

Quality loss with each conversion

Best Practices

- Plan format requirements
- Document conversion steps
- Preserve original files

The "Annotation Consistency" Problem

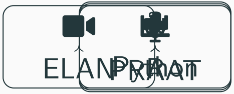


Solution

Create annotation guidelines with:

- Standard labels
- Clear examples
- Version control

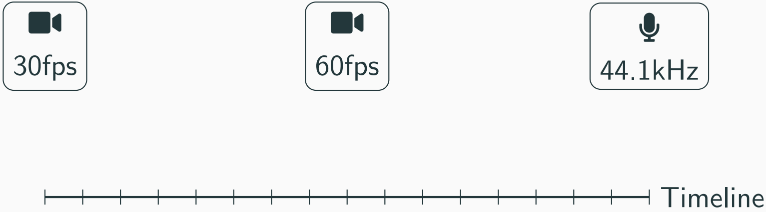
The "Multiple Tool Chain" Challenge



Solution

- Document data transformations
- Create automated pipelines
- Use version control

The "Frame Rate Mismatch" Problem



Issues

- Temporal misalignment
- Sync drift
- Different sample rates

Solutions

- Use sync markers
- Standard recording setup
- Time code generators

Testing and Validation

Automated Tests

- Unit tests
- Integration tests
- Continuous Integration

Manual Validation

- Code review
- Documentation review
- Reproducibility check

Best Practices Checklist

Setup

- ✓ Use version control
- ✓ Create virtual environment
- ✓ Document dependencies

Maintenance

- ✓ Regular testing
- ✓ Update documentation
- ✓ Backup data

Remember
Reproducibility is a journey, not a destination

Resources

- The Turing Way: `https://the-turing-way.netlify.app`
- Git Guide: `https://guides.github.com`
- Python Packaging: `https://packaging.python.org`
- Docker Docs: `https://docs.docker.com`

Questions?

Ethical Data Management

Privacy Protection

- Data Anonymization
- Consent Management
- Access Controls
- Data Minimization

Security Measures

- Encryption
- Secure Storage
- Access Logging
- Deletion Protocols

Ethical Data Management: Real-World Scenarios

Scenario 1: Interview Research

Context: Studying sensitive community experiences

- Collect narratives about marginalized groups
- Participants share deeply personal stories
- **Ethical Challenges:**
  - Protecting individual identities
  - Preventing potential harm
  - Maintaining trust

Scenario 2: Large-Scale Surveys

Context: Digital health research

- Collect sensitive medical information
- Anonymous online questionnaires
- **Ethical Challenges:**
  - Data anonymization
  - Informed consent
  - Secure data storage

Ethical Considerations: Practical Approaches

Practical Mitigation Strategies

Data Protection:

- Anonymization techniques
- Pseudonymization
- Encryption
- Access controls

Participant Safety:

- Clear consent forms
- Option to review/delete data
- Transparent data use
- Minimal personal information

Ethical Considerations: Practical Approaches

Key Ethical Principles

- Informed Consent
- Data Minimization
- Purpose Limitation
- Confidentiality
- Right to Withdraw

Ethical Dilemmas: Interactive Discussion

Ethical Decision-Making

"You've collected video data of children's language development. How do you balance research value with participant privacy?"

Potential Considerations

- Parental consent
- Video masking techniques
- Long-term data storage
- Future use restrictions
- Right to be forgotten

Global Perspectives on Research Ethics

Regulatory Frameworks

- GDPR (European Union)
- IRB Protocols
- Institutional Guidelines
- International Research Standards

Institutional Responsibilities

- Ethics review boards
- Training programs
- Compliance monitoring
- Ongoing ethical education

Critical

Ethical considerations are not an afterthought but a fundamental aspect of responsible research

Identity-Preserving Video Processing with MediaPipe

Core Concept
A framework for processing videos while preserving privacy and extracting movement data.

Input Processing

- Single-person video input
- Frame-by-frame MediaPipe analysis
- Body silhouette segmentation

Output Generation

- Masked video with preserved background
- Overlaid kinematic tracking
- Time-series data extraction

Key Features

- Tracks 33 body landmarks
- Tracks 42 hand landmarks
- Tracks 478 facial mesh points

Masking Techniques

Today's Focus

- Introduction to Data Privacy
- Video Masking Strategies
- Audio Anonymization
- Ethical Considerations
- Hands-on Masking Techniques

Schedule

1. Q&A (5 min)
2. Technical Session: Masking 101 (15 min)
3. Collaborative Discussion (5 min)

Why Masking Matters

Privacy Concerns

- Protect individual identities
- Comply with data protection regulations
- Prevent unauthorized identification
- Ethical research practices

Research Utility

- Preserve meaningful data
- Maintain research insights
- Allow data sharing
- Support reproducibility

Video Masking Strategies: Overview

- **Two Primary Approaches:**
  1. **Hiding Strategy:** Complete de-identification
  2. **Masking Strategy:** Preserve useful information
- **Hiding Techniques Include:**
  - Blackout
  - Gaussian Blur
  - Contour Preservation
  - Video Inpainting
- **Masking Techniques Include:**
  - Skeleton Overlay
  - Face Mesh Preservation
  - Face Swapping
  - 3D Avatar Rendering

Hiding Techniques Detailed

Hiding Methods

- **Blackout:** Complete obscuration
- **Gaussian Blur:** Soft obscuration
- **Contour Preservation:** Maintain shape
- **Video Inpainting:** Background reconstruction

Masking Techniques

Video Masking

- Face De-identification
- Background Removal
- Feature Preservation
- Quality Control

Trade-offs

- Privacy vs. Utility
- Processing Speed
- Storage Requirements
- Analysis Impact

Practical Implementation

Tools

- EnvisionBOX Platform
- MaskAyone
- MediaPipe
- Version Control
- Documentation Tools

Workflow

- Data Collection
- Processing Pipeline
- Quality Checks
- Archive Preparation

Research Roadmap

Planning Steps

1. Define Research Goals
2. Identify Data Requirements
3. Plan Processing Pipeline (Local vs Server)
4. Consider Ethics & Privacy
5. Implement Solutions

Quiz Preparation/Think about

- Open Science Impact and Ethical Considerations
- Technical Requirements and Practical Implementation

Resources & Support

Useful Links

- EnvisionBOX Guide
- Jupyter Documentation
- GitHub Repository
- Tutorial Videos

Help Available

- Direct support
- Online resources
- Troubleshooting guide
- Community forum